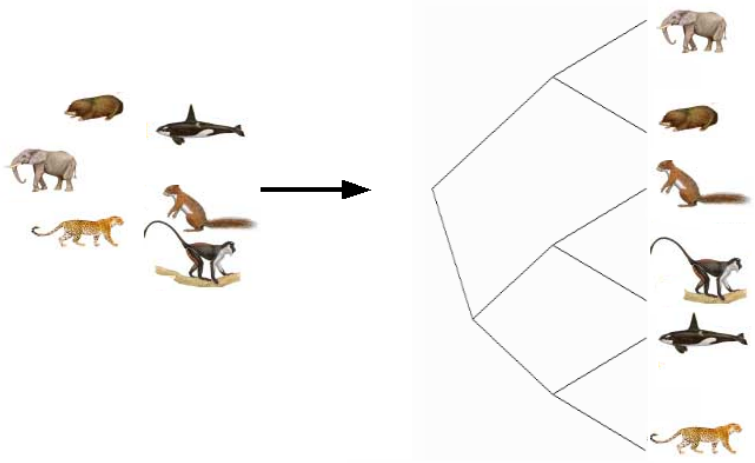
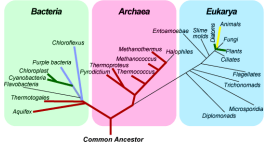


Introduction à la phylogénie moléculaire

Thomas Bigot

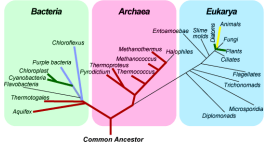
thomas.bigot@univ-lyon1.fr

18 mars 2015



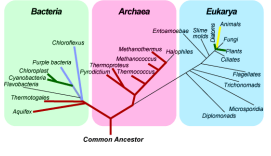
Première partie I

Introduction



1 - Introduction

1. Qu'est-ce que la phylogénie



La phylogénie
Concept d'homologie

La phylogénie

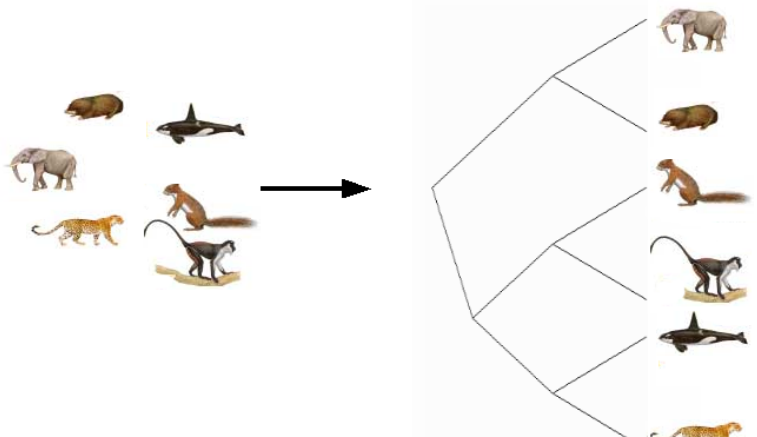
Étude des relations évolutives au sein d'un groupe d'entités homologues (= provenant d'un ancêtre commun).

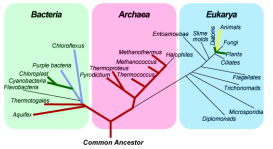
Le concept d'homologie :

Les données observées aujourd'hui en biologie (molécules, organes, structures...) sont le produit de l'évolution depuis 3,5-4Mds d'années.

Étude de l'évolution des organismes : comparaison de certaines caractéristiques **que leur ancêtre commun possédait**.

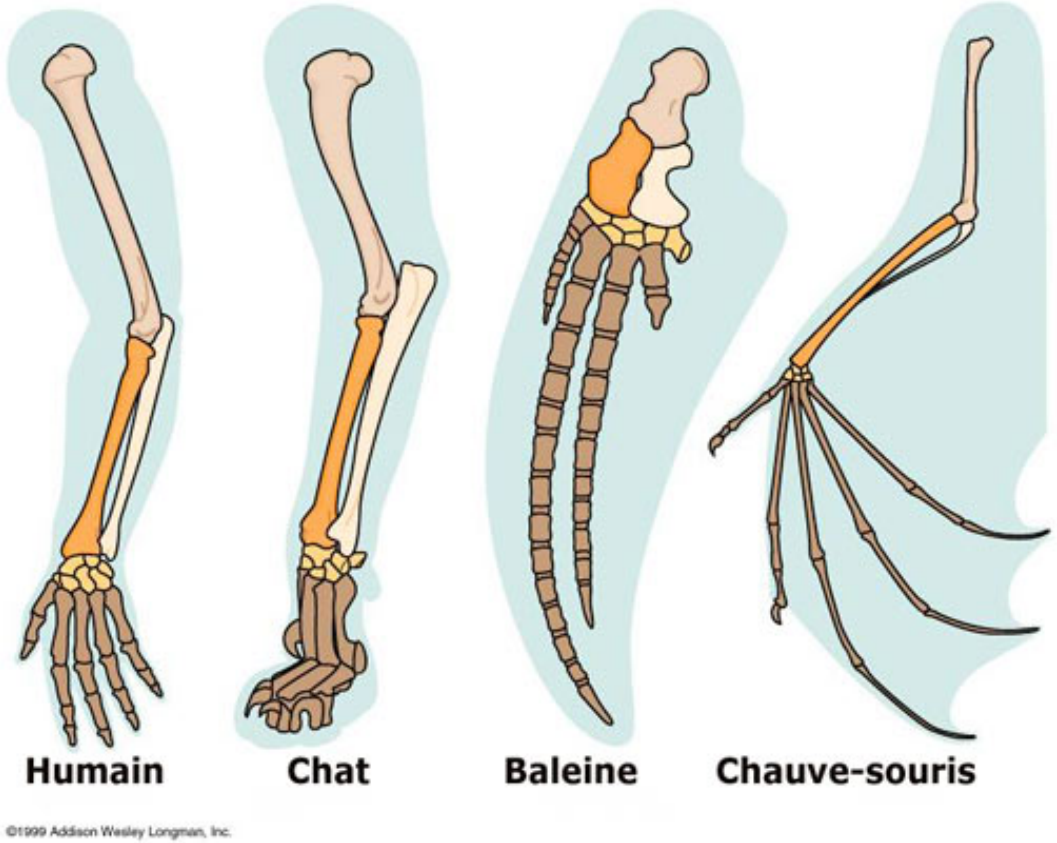
Ces caractéristiques sont dites homologues.

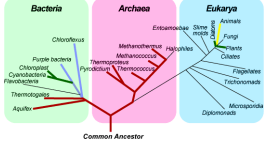




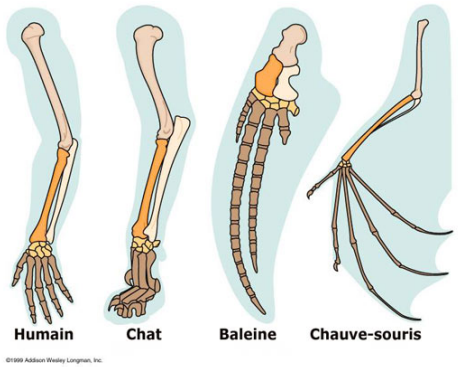
Exemple d'homologie

Évolution des mammifères :
Exemple de caractère homologue : le membre antérieur





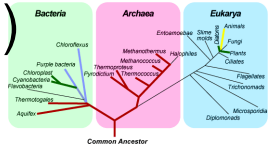
Exemple d'homologie



Quels arguments pour l'homologie d'organes / structures ?

- la position dans le corps
- l'organisation interne (os, muscles)
- le développement
- l'expression des gènes

L'explication la plus plausible pour expliquer tous ces points communs est qu'**ils ont été hérités d'un ancêtre commun** (=homologues) plutôt qu'inventés indépendamment 4 fois.

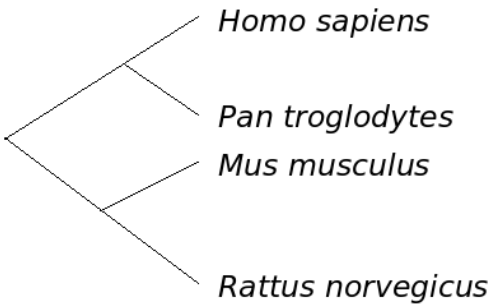


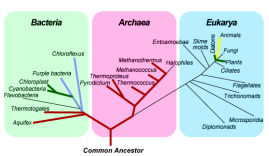
Comment faire de la phylogénie ?

- Choix des espèces
- Choix des caractères HOMOLOGUES

Exemple : caractères homologues : pouce et queue.

Espèce	Pouce opposable	Queue
Homo sapiens	oui	bourgeon caudal
Pan troglodytes	oui	bourgeon caudal
Mus musculus	non	oui
Rattus norvegicus	non	oui



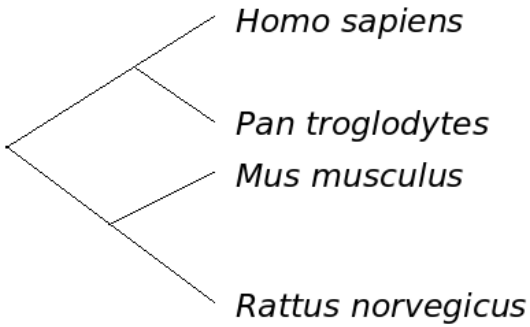


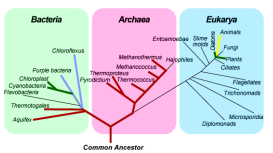
Phylogénie moléculaire

Phylogénie moléculaire :
comparaison de séquences homologues d'ADN ou de protéines pour l'inférence de l'histoire évolutive des séquences.

Si l'on fait **l'hypothèse que l'histoire des séquences est la même que celle des organismes** qui les contiennent, on a ainsi accès à la phylogénie des organismes, populations, souches, espèces...

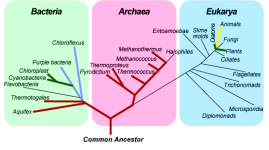
Espèce	Séquence
Homo sapiens	ATTGCCCTGA...
Pan troglodytes	ATAGCCCTGA...
Mus musculus	AGTGCCCTGA...
Rattus norvegicus	AGTGCACTGA...





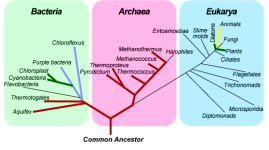
Deuxième partie II

Des séquences aux arbres : les étapes



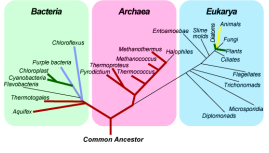
Principe d'une analyse de phylogénie moléculaire

1. **Rassembler des séquences homologues**
   On veut comparer des choses comparables : qui proviennent d'une séquence ancestrale commune
2. **Aligner ces séquences**
   De même, on veut comparer des sites qui proviennent du même site ancestral
3. **Reconstruire l'arbre**
   Différentes méthodes peuvent être adaptées à différents types de données
4. **Raciner et interpréter l'arbre**
   L'interprétation dépendra notamment des indices de confiance associées aux branches...



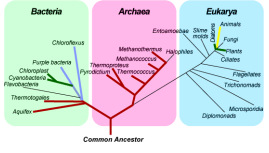
2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
   - Nature des séquences utilisées
   - Altérations des séquences
   - Construction d'un jeu de données
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre



2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
   - Nature des séquences utilisées
   - Altérations des séquences
   - Construction d'un jeu de données
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre



Ensemble de séquences homologues

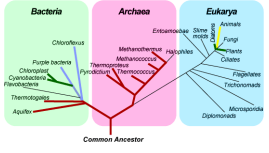
Nature des séquences utilisées

- ADN codant ou non codant
- Protéines
- Mitochondrial, chloroplastique, viral, nucléaire, plasmidique...

Le marqueur choisi dépend de la question biologique posée.
Vitesse d'évolution des

- séquences de virus à ARN
- > séquences de virus à ADN
- > séquences mitochondriales
- > séquences nucléaires.
  - ADN non codant (moins contraint)
  - > ADN codant
  - > Protéines (redondance du code)

En outre, certains gènes évoluent plus vite que d'autres...



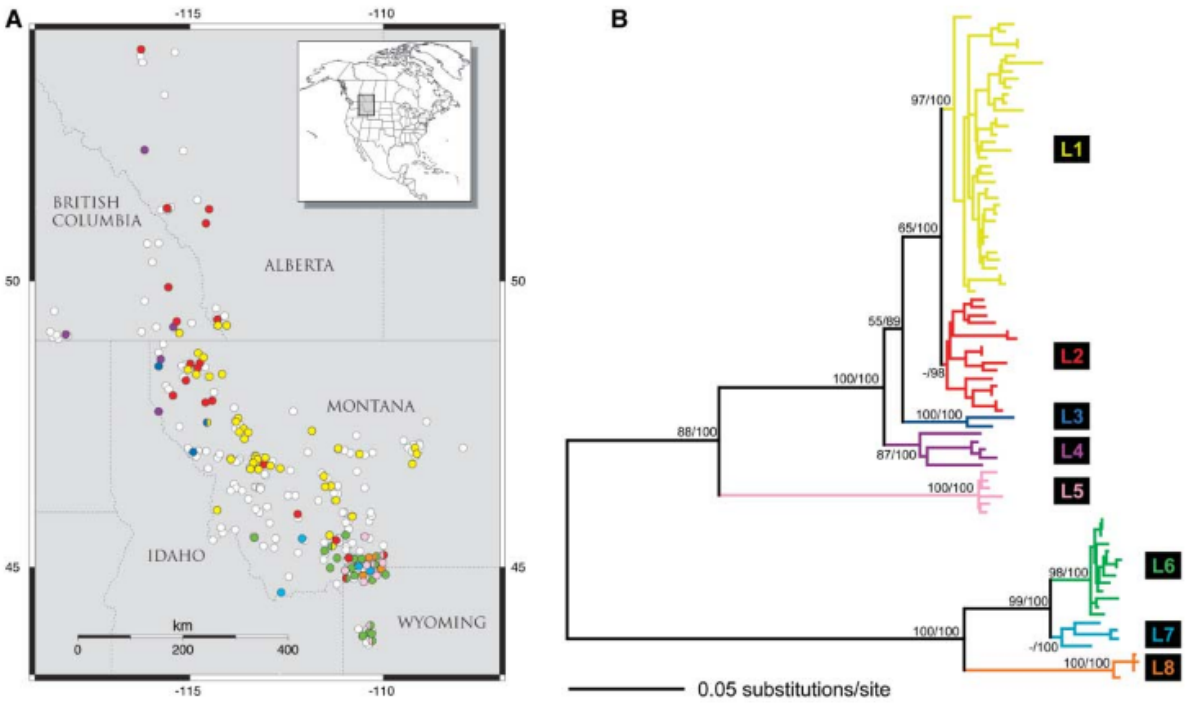
Vitesse d'évolution

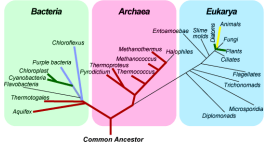
Suivi de populations

A virus reveals population structure and recent demographic history of its carnivore host.

Biek *et al.*, *Science*, 2006

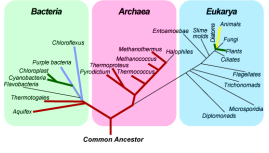
. . . ou comment l'analyse phylogénétique d'un virus nous permet de comprendre l'évolution des populations de Puma





2 - Des séquences aux arbres : les étapes

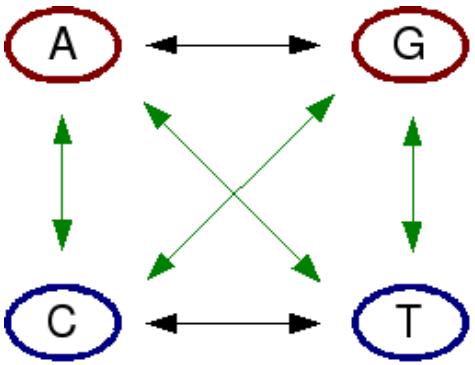
1. Rassembler séquences homologues
   - Nature des séquences utilisées
   - Altérations des séquences
   - Construction d'un jeu de données
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre

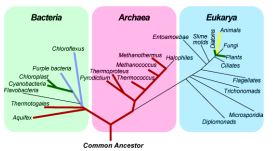


Ensemble de séquences homologues

Altérations des séquences

- Substitution
  - Transition
  - Transvertion
- Insertion/délétion

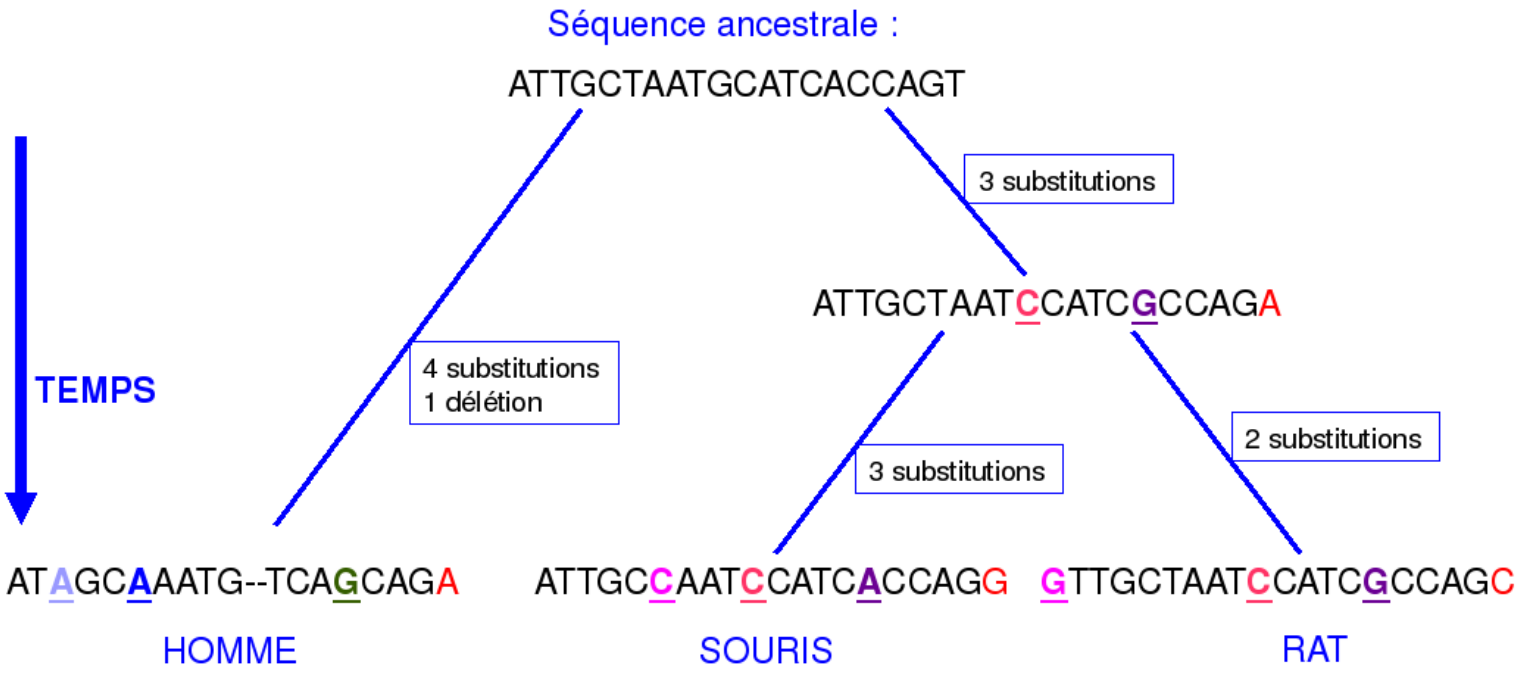




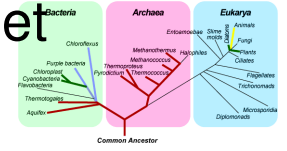
Ensemble de séquences homologues

Altérations des séquences

(Ceci n'est pas un arbre phylogénétique.)

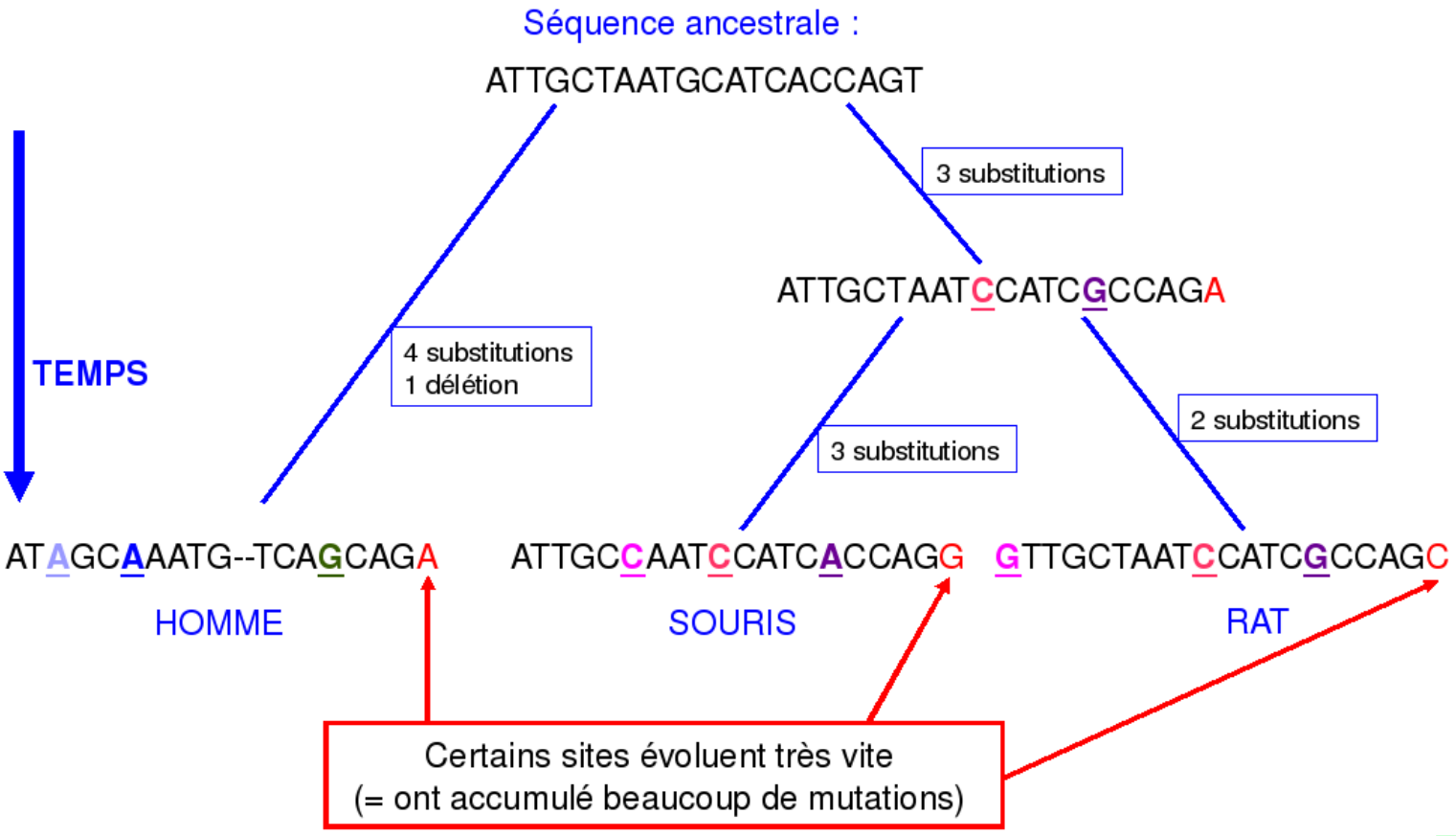


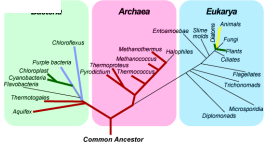
Si on veut reconstruire l'histoire évolutive des séquences actuelles, il faut d'abord les aligner pour trouver les positions des insertions et délétions



Ensemble de séquences homologues

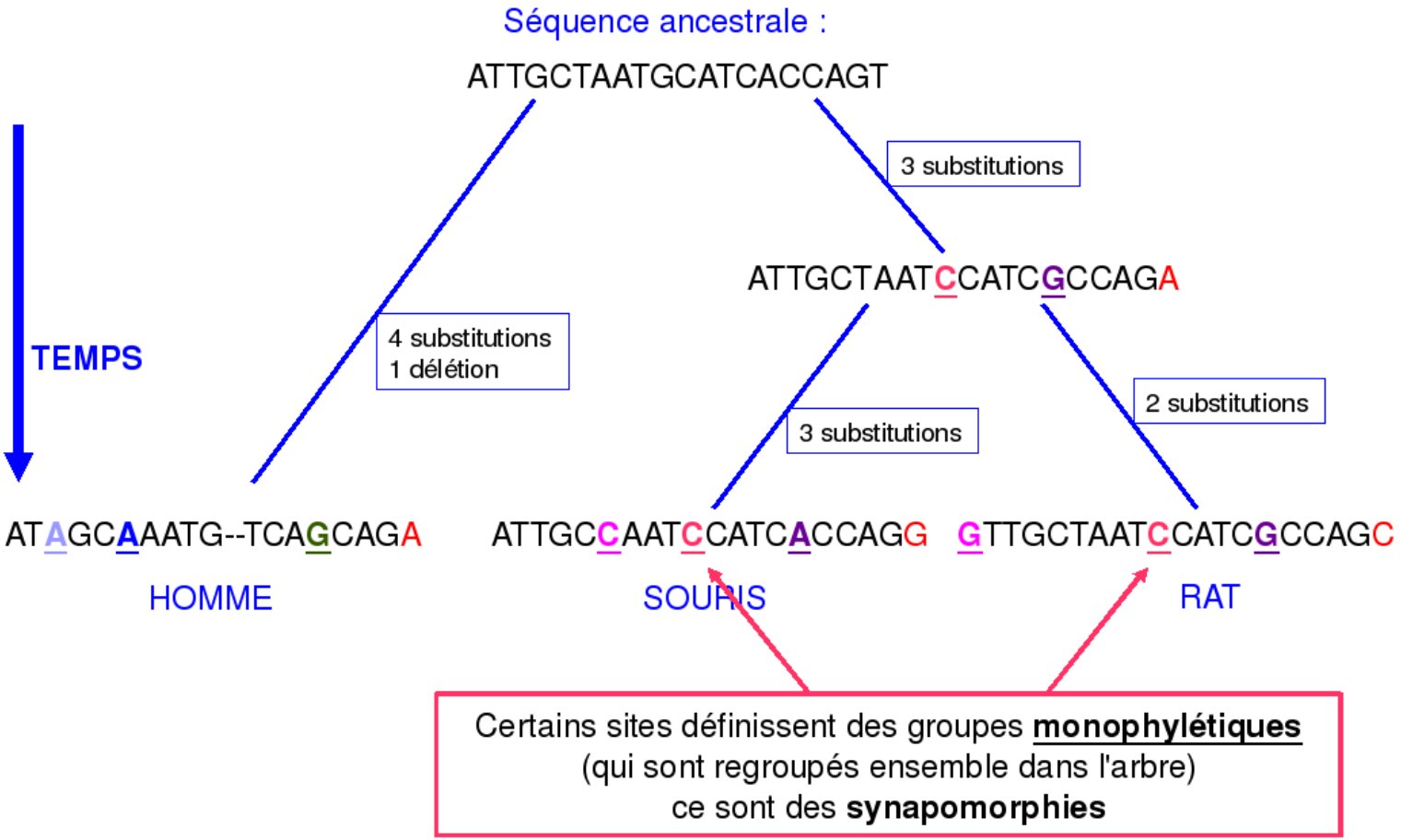
Altérations des séquences

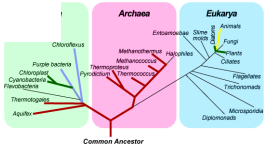




Ensemble de séquences homologues

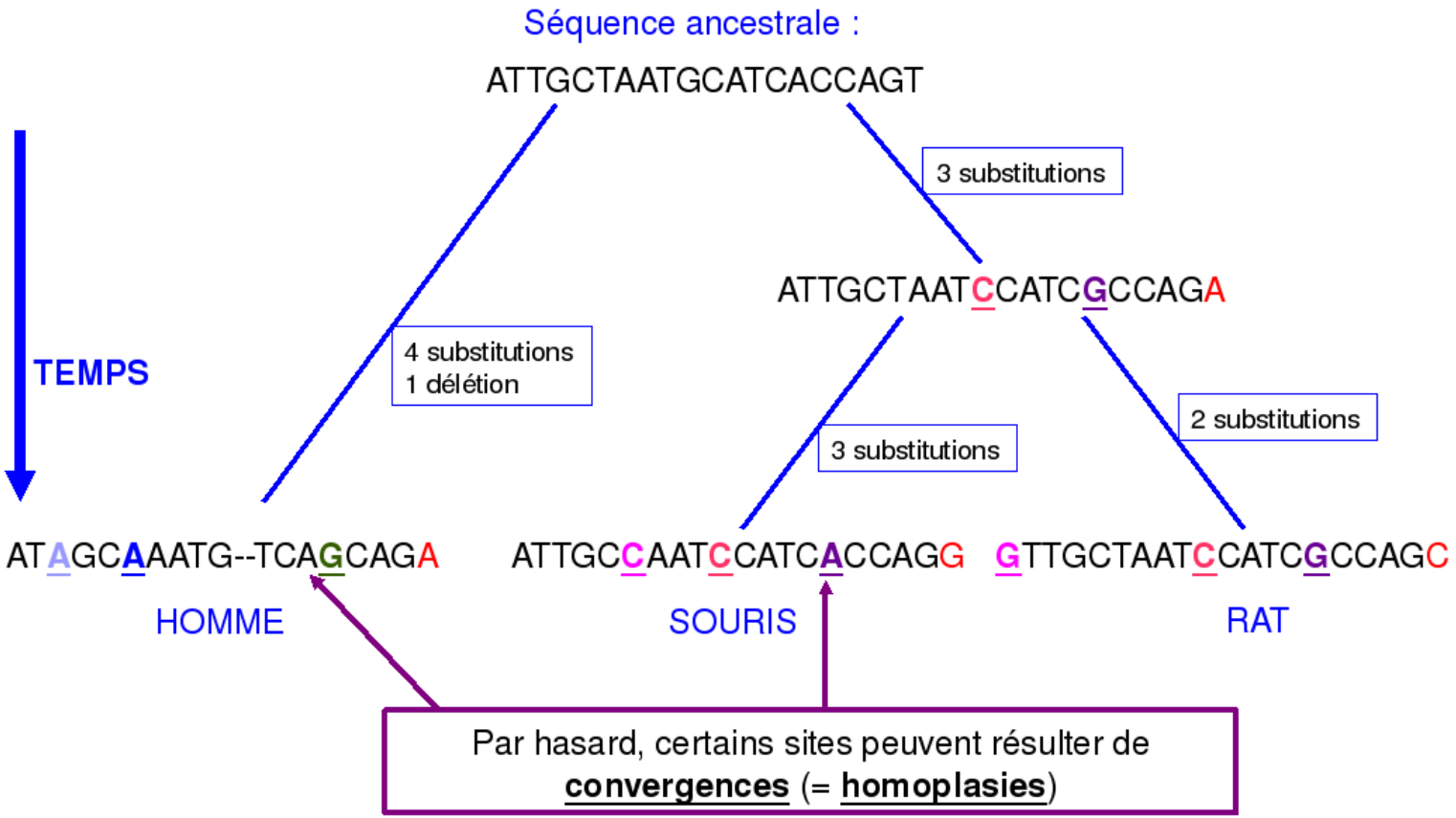
Altérations des séquences

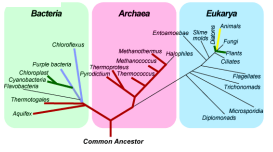




Ensemble de séquences homologues

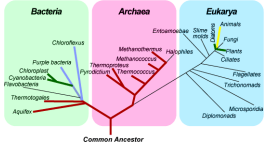
Altérations des séquences





2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
   - Nature des séquences utilisées
   - Altérations des séquences
   - Construction d'un jeu de données
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre



Ensemble de séquences homologues

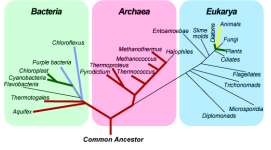
Construction d'un jeu de données

Hypothèse

Deux séquences sont homologues si elles sont tellement similaires qu'une convergence semble très peu probable.

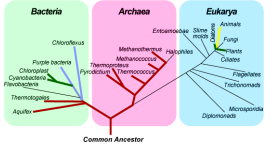
Utilisation de logiciels de recherche d'homologie (ex : Blast)
Utilisation de banques de familles de gènes homologues : ex :

- Hovergen (gènes de Chordés de Swissprot)
- HomolENS (génomes complets d'Eucaryotes d'Ensembl)
- HOGENOM (génomes complets)
- COGs (génomes complets de procaryotes)...



2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre



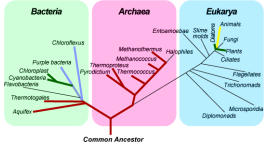
Alignements

- Principe :
  Comparer des séquences issues d'organismes différents pour aligner les sites homologues

Hypothèse

le "meilleur" alignement est celui qui suppose le moins d'événements évolutifs

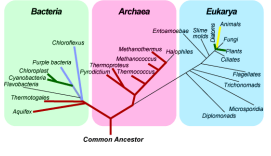
- Utilisation d'algorithmes d'alignement de séquences (e.g. Clustalw, Muscle, T-Coffee...) :
  - on attribue des pénalités aux substitutions (différentes pénalités pour différentes substitutions)
  - on attribue des pénalités à l'insertion ou à l'extension d'une brèche
  - l'algorithme recherche l'alignement qui a le score de pénalités le plus faible.



Alignements

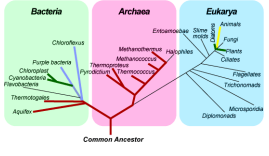
- Chaque colonne de l'alignement (site) est supposée contenir des résidus (nucléotides, acides aminés) homologues qui peuvent avoir subit des changements.
- **Il faut toujours vérifier l'alignement à la main, et supprimer les sites dont l'homologie est incertaine !**

```
Acrop_ten  RKENPLLSPV NGLLVDLPSP SNISYLWNFG SLLGLCLAMQ IVTGCFLSMH YCAEVGLAFA SVGQ-NSDVN
Boa_con    QKILMLFGL- ------LPVA TNISTWWNFG SMLLTCSMIQ VLTGFFLAVH YTANINLAFS SIVHIMRDVP
Homo_sap   RKINPLMKLI NHSFIDLPTP SNISAWWNFG SLLGACLILQ ITTGLFLAMH YSPDASTAFS SIAHITRDVN
Gallu_gal  RKSHPLLKMI NNSLIDLPAP SNISAWWNFG SLLAVCLMTQ ILTGLLLAMH YTADTSLAFS SVAHTCRNVQ
Testu_gra  RKTHPMMKII NNSFIDLPSP SNISAWWNFG SLLGICLILQ IITGIFLAMH YSPNISLAFS SVAHITRDVQ
Allig_mis  RKSHPIIKLI NRSLIDLPTP SNISAWWNFG SLLGLTLLIQ ILTGFFLMMH FSSSDTLAFS SVSYTSREVW
Gekko_gec  RKHETLLKII NHSLIDLPTP SNISTWWNFG SLLGLCLILQ IITGLFLSMH YTANTSLAFQ SLTHVIRDVH
Squal_aca  RKTHPLIKIV NHALVDLPSP SNISIWWNFG SLLGLCLIIQ ILTGLFLAMH YTADISTAFS SVVHICRDVN
Xenop_lae  RKSHPLIKII NNSFIDLPTP SNISSLWNFG SLLGVCLIAQ IITGLFLAMH YTADTSMAFS SVAHICFDVN
```



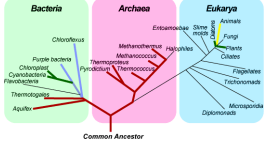
2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
   - Recherche du meilleur arbre
   - Les différents méthodes de reconstruction
   - Les hypothèses de reconstruction phylogénétique
   - Description des méthodes de reconstruction
   - Comparaison des méthodes
4. Racinement et interprétation de l'arbre



2 - Des séquences aux arbres : les étapes

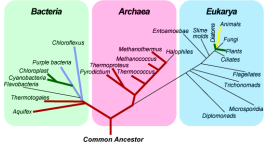
1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
   - Recherche du meilleur arbre
   - Les différents méthodes de reconstruction
   - Les hypothèses de reconstruction phylogénétique
   - Description des méthodes de reconstruction
     - Maximum de parcimonie
     - Minimum d'évolution
     - Maximum de vraisemblance
   - Comparaison des méthodes
4. Racinement et interprétation de l'arbre



Méthodes de reconstruction d'arbres

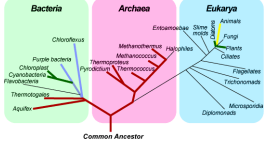
Recherche du *meilleur* arbre selon un critère donné

- Meilleure méthode :
  essayer toutes les topologies possibles, et choisir la meilleure respectivement au critère choisi (vraisemblance, longueur, nombre de pas de parcimonie...)
- Problème :
  le nombre de topologies possibles augmente factoriellement avec le nombre de séquences :
  - 5 séquences : 15 topologies
  - 10 séquences : 202 725 topologies
  - 20 séquences : 221 643 095 476 699 771 875 topologies
- On a donc recours à des heuristiques



2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
   - Recherche du meilleur arbre
   - Les différents méthodes de reconstruction
   - Les hypothèses de reconstruction phylogénétique
   - Description des méthodes de reconstruction
     - Maximum de parcimonie
     - Minimum d'évolution
     - Maximum de vraisemblance
   - Comparaison des méthodes
4. Racinement et interprétation de l'arbre



Méthodes de reconstruction d'arbres

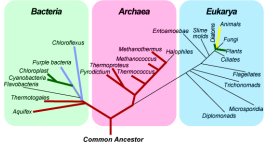
Les différents méthodes de reconstruction des arbres

Méthode du maximum de parcimonie on cherche l'arbre tel que le nombre d'événements de substitution est minimal

Méthode du minimum d'évolution on cherche l'arbre tel que sa longueur est minimale

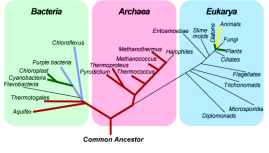
Méthode du maximum de vraisemblance on cherche l'arbre qui maximise la probabilité d'occurrence des données

Méthodes bayésiennes on recherche l'arbre le plus probable sachant les données



2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
   - Recherche du meilleur arbre
   - Les différents méthodes de reconstruction
   - Les hypothèses de reconstruction phylogénétique
   - Description des méthodes de reconstruction
     - Maximum de parcimonie
     - Minimum d'évolution
     - Maximum de vraisemblance
   - Comparaison des méthodes
4. Racinement et interprétation de l'arbre

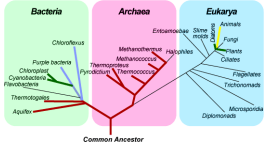


Méthodes de reconstruction d'arbres

Hypothèses de reconstruction phylogénétique

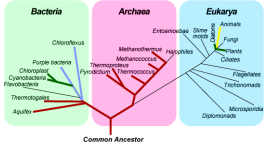
Hypothèses communes

- Utilisation des événements de substitutions (pas les insertions-délétions) pour reconstruire l'histoire de séquences
- Chacun des sites évolue **indépendamment** des autres sites
- Chaque site évolue à une vitesse constante sur l'arbre (excepté parcimonie)
- Le modèle d'évolution est le même en chacun des sites de l'alignement et sur l'ensemble de l'arbre



2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
   - Recherche du meilleur arbre
   - Les différents méthodes de reconstruction
   - Les hypothèses de reconstruction phylogénétique
   - Description des méthodes de reconstruction
     - Maximum de parcimonie
     - Minimum d'évolution
     - Maximum de vraisemblance
   - Comparaison des méthodes
4. Racinement et interprétation de l'arbre



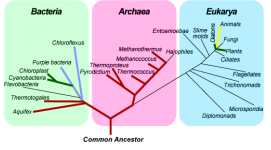
Maximum de parcimonie

Principe

Rasoir d'Ockham

L'hypothèse qui nécessite le moins d'événements est favorisée

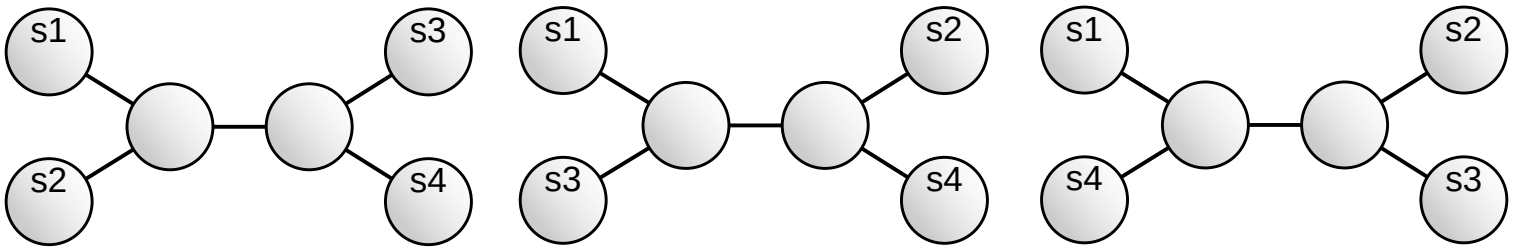
- Principe associé au rasoir d'Ockham
- Toutes les mutations ont le même poids
- Heuristique : essayer un grand nombre d'arbres et choisir celui (ou ceux !) qui minimisent le nombre d'événements de substitutions : plusieurs arbres peuvent être équi-parcimonieux



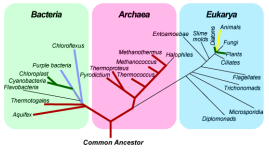
Maximum de parcimonie

Algorithme

```
          Site X  Site Y
s1: TAGCAGTAAGCCT---GGA
s2: TAGCAGTACGCCTACCGGA
s3: TAGCCGTAGGCCTACCGGA
s4: TAGCCGTAAGCCTACCGGA
```



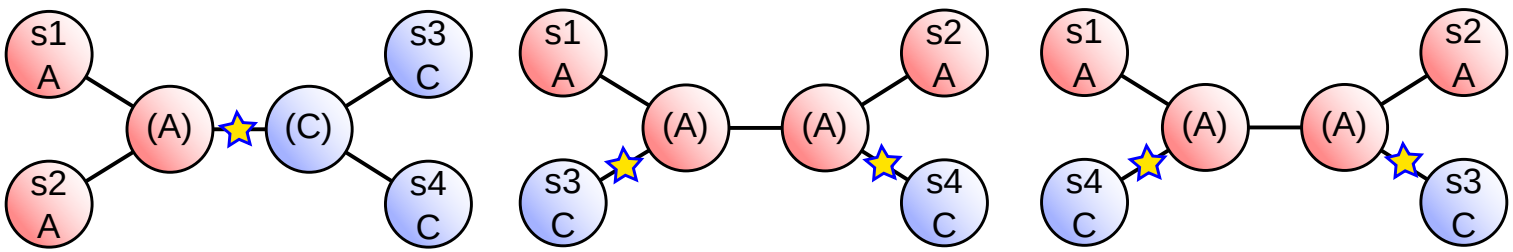
Adapté de Bigot, 2013



Maximum de parcimonie

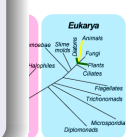
Minimisation du nombre d'évènements sur **Site X**

```
          Site X  Site Y
s1:  TAGCAGTAAGCCT---GGA
s2:  TAGCAGTACGCCTACCGGA
s3:  TAGCCGTAGGCCTACCGGA
s4:  TAGCCGTAAGCCTACCGGA
```



Séquences Ancestrales

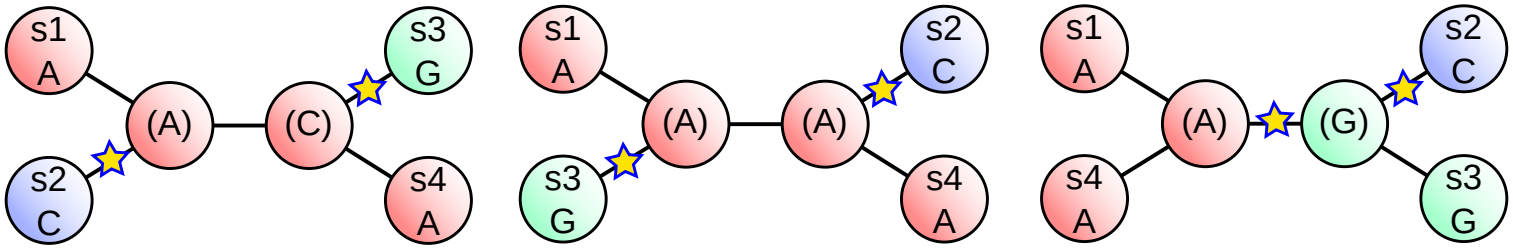
Les séquences ancestrales ne sont pas définies de manière unique, plusieurs possibilités sont également parcimonieuses.



Maximum de parcimonie

Minimisation du nombre d'évènements sur un 2ème site

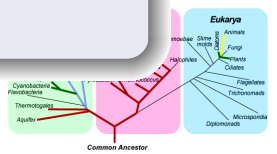
```
          Site X  Site Y
s1: TAGCAGTAAGCCT---GGA
s2: TAGCAGTACGCCTACCGGA
s3: TAGCCGTAGGCCTACCGGA
s4: TAGCCGTAAGCCTACCGGA
```



Site informatif

Un site est dit *informatif* si et seulement si il y a au moins 2 nucléotides différents chacun répété au moins 2 fois.

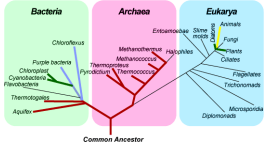
Pour le site `Y`, Tous les arbres requièrent le même nombre de changement donc ce site n'est pas informatif



Maximum de parcimonie

Caractéristiques

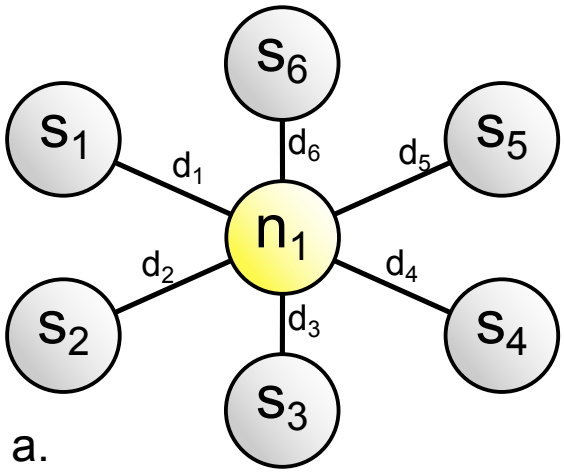
- Avantages
  Applicable à des caractères généraux (phénotypiques, moléculaires, binaires, ...)
- Inconvénients
  - Longs calculs si beaucoup de séquence car beaucoup d'arbres à tester, si trop de séquences on ne teste pas tous les arbres qui existent heuristique de recherche parmi tous les arbres possibles
  - Il peut y avoir des exaequos
  - Suppose une vitesse d'évolution constante
  - La position des changements sur chaque branche n'est pas unique donc pas de longueurs de branches.



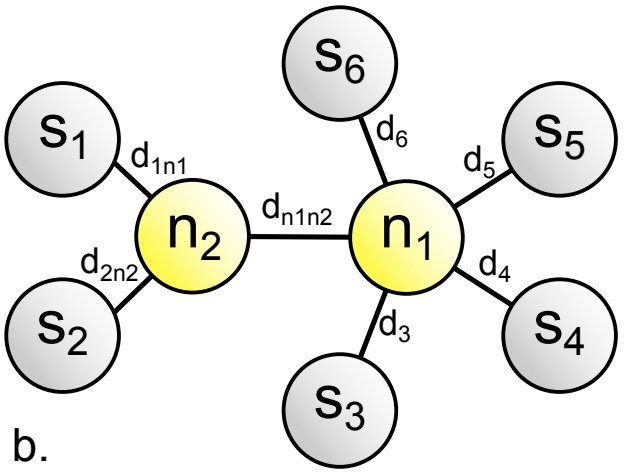
Minimum d'évolution (méthodes des distances)

Principe

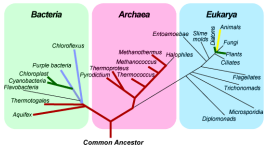
- Apparenté lui aussi au rasoir d'Ockham
- Basé sur un **modèle évolutif** : différentes substitutions peuvent avoir différentes probabilités
- **Heuristique** : *Neighbor Joining*
  On calcule les distances deux à deux entre séquences ; on part d'une topologie en étoile, puis les deux séquences les moins distantes sont regroupées, puis la séquence la moins distante de ces deux séquences, etc...



a.



b.



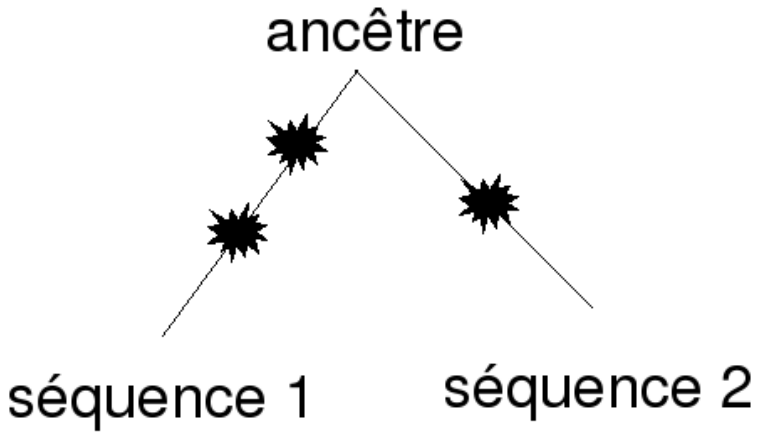
Minimum d'évolution (méthodes des distances)

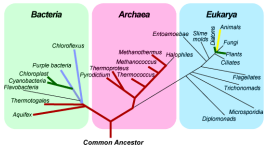
Calcul des distances évolutives

Distance évolutive

Nombre de substitutions estimées produites depuis la séparation de l'ancêtre commun, sur les deux lignées divisé par le nombre de sites. Une distance est donc exprimée en substitutions estimées par sites

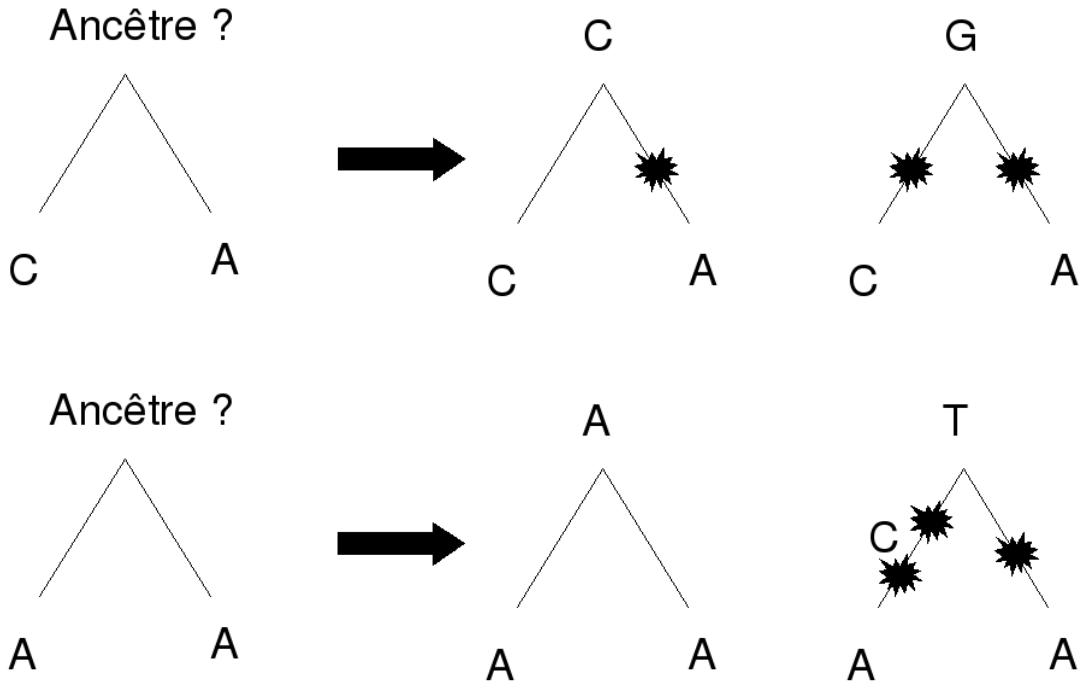
L'estimation se fait au travers d'un modèle d'évolution



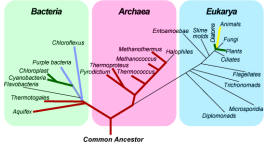


Minimum d'évolution (méthodes des distances)

Substitutions cachées



Comment inférer le nombre réel de substitutions produites ?

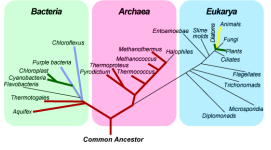


Minimum d'évolution (méthodes des distances)

Modèles d'évolution

Utilisation de modèles d'évolution, en faisant des hypothèses sur la nature des processus évolutif.
Un modèle permet d'essayer de tenir compte des substitutions cachées

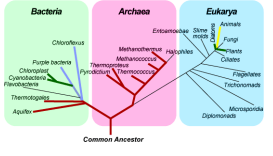
- Modèle de Jukes et Cantor (1969)
  Toutes les substitutions sont équiprobables
- Modèle de Kimura à 2 paramètres
  Il existe deux types de substitutions (transitions / transversions) qui ont des probabilités différentes.
- Il en existe bien d'autres, qui modélisent plus finement le processus d'évolution des séquences



Maximum de vraisemblance
Principe

- Principe d'inférence statistique largement utilisé
- Pas d'hypothèse de rasoir d'Ockham
- Basé sur un modèle évolutif : différentes substitutions peuvent avoir différentes probabilités
- Permet de modéliser plus finement l'évolution des séquences que les méthodes de distance
- Heuristique : comme pour le maximum de parcimonie, mais un seul arbre est toujours renvoyé

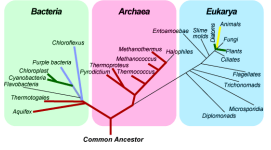
Pour calculer la vraisemblance d'un arbre, on somme les probabilités de l'ensemble des scénarios possibles.
On garde l'arbre ayant obtenu la meilleure vraisemblance.



Maximum de vraisemblance
Caractéristiques

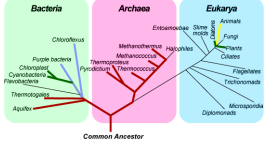
- Avantages
  Méthode la mieux justifiée d'un point de vue théorique
  Cette méthode marche mieux que toutes les autres dans la pluspart des cas

- Inconvénients
  Quand trop de séquences le nombre d'arbres à tester est trop grand, on explore alors qu'une partie de l'ensemble des arbres possibles (heuristique de recherche parmi tout les arbres possibles). On perd alors la certitude mathématique d'avoir trouvé l'arbre le plus vraissemblable.



2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
   - Recherche du meilleur arbre
   - Les différents méthodes de reconstruction
   - Les hypothèses de reconstruction phylogénétique
   - Description des méthodes de reconstruction
     - Maximum de parcimonie
     - Minimum d'évolution
     - Maximum de vraisemblance
   - Comparaison des méthodes
4. Racinement et interprétation de l'arbre

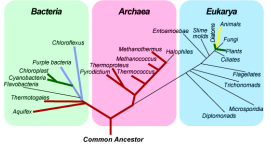


Comparaison des méthodes

Temps d'éxecution

TABLE 1. Average run times for various methods. The computing times were measured on a 1.8-GHz (1 Go RAM) PC with Linux. For PHYML, the number in parentheses is the average number of refinement stages.

	Simulations		Real data	
Method	40 taxa (500 bp)	100 taxa (500 bp)	218 taxa (4,182 bp)	500 taxa (1,428 bp)
DNADIST+ NJ/BIONJ	0.3 sec	2.3 sec	50 sec	2 min, 19 sec
DNADIST+ Weighbor	1.5 sec	22 sec	4 min, 52 sec	58 min, 40 sec
DNAPARS	0.5 sec	6 sec	4 min, 4 sec	13 min, 12 sec
PAUP*	3 min, 21 sec	1 hr, 4 min		
PAUP*+ NJ	1 min, 10 sec	22 min	10 hr, 50 min	
MrBayes	2 min, 6 sec	32 min, 37 sec		
fastDNAml	1 min, 13 sec	26 min, 31 sec		
NJML	15 sec	6 min, 4 sec		
MetaPIGA	21 sec	3 min, 27 sec	4 hr, 45 min	9 hr, 4 min
MetaPIGA+ NJ	6 sec	23 sec	1 hr, 40 min	3 hr
PHYML	2.7 sec (6.4)	12 sec (8.3)	8 min, 13 sec (15)	11 min, 59 sec (13)

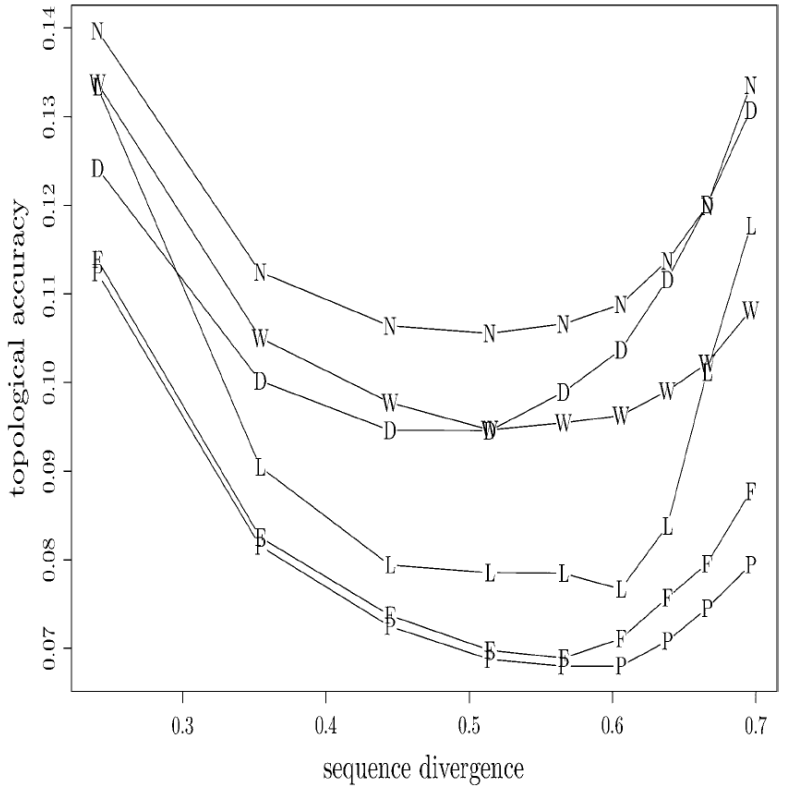


Comparaison des méthodes

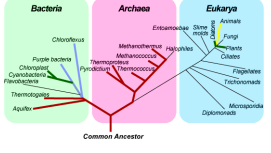
Comparaison des méthodes sur des données simulées

Précisions

- N=Neighbor Joining (distances)
- W=Weighbor (distance)
- D=DNAPARS (parcimonie)
- L=NJML (max vraisemblance)
- F=fastDNAml (max vraisemblance)
- P=PHYML (max vraisemblance)

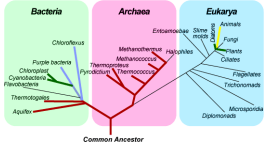


Guindon et Gascuel, Syst. Biol. 2003.



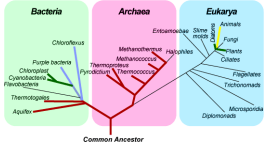
2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre
   - Raciner un arbre
   - Indices de confiance
   - Difficultés d'interprétation



2 - Des séquences aux arbres : les étapes

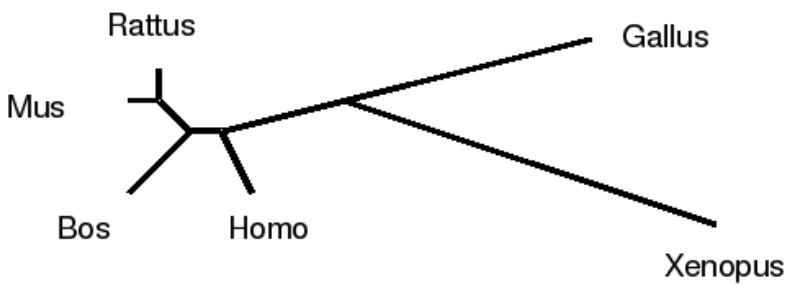
1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre
   - Raciner un arbre
   - Indices de confiance
   - Difficultés d'interprétation
     - Notions d'orthologie et paralogie
     - Différentes vitesses d'évolution
     - Biais compositionels
     - Phénomène de coalescence



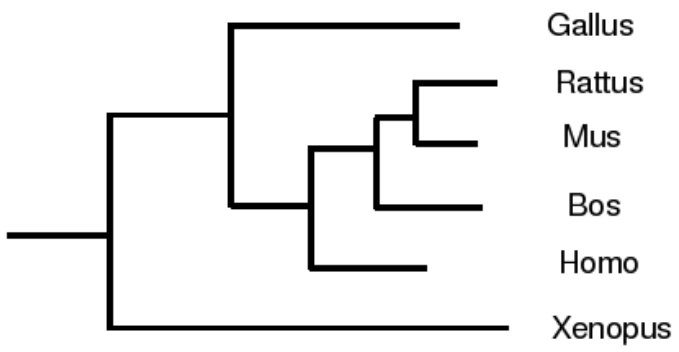
Analyse de l'arbre obtenu

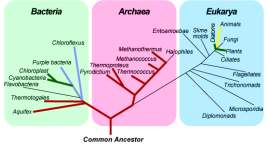
Raciner un arbre

Arbre non raciné



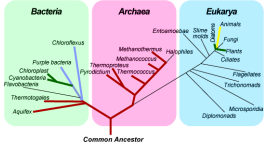
Arbre raciné





2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre
   - Raciner un arbre
   - Indices de confiance
   - Difficultés d'interprétation
     - Notions d'orthologie et paralogie
     - Différentes vitesses d'évolution
     - Biais compositionels
     - Phénomène de coalescence

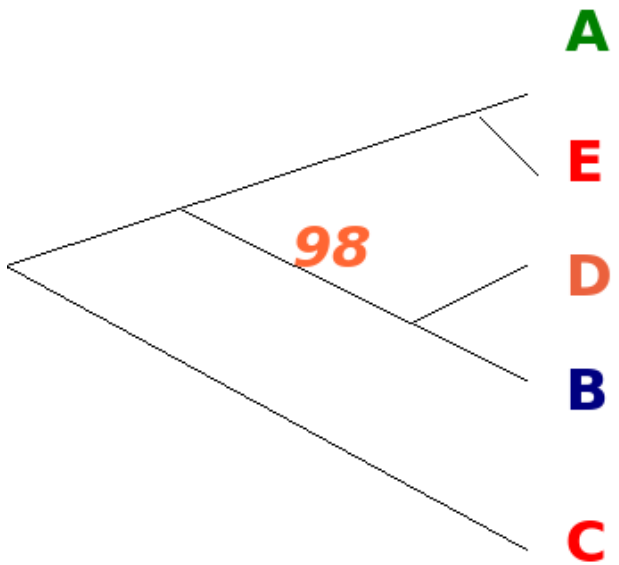


Analyse de l'arbre obtenu

Indices de confiance

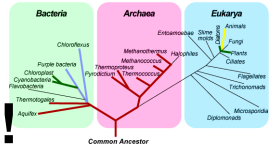
Permettent de savoir à quel point on peut se fier à un branchement particulier.

Les séquences (D et B) sont séparées des séquences (A, E et C) avec un indice de confiance de 98 %.



Bootstrap = mesure du soutien de l'arbre par les données

- Faible bootstrap à une branche : peu de signal contenu dans les données pour cette branche, d'après la méthode
- **N'aide pas à la sélection de la méthode phylogénétique !**



Indices de confiance

Bootstrap - technique

Alignement réel

```
ACGTACATAGTATAGCGTCTAGTGGTACCGTATG
AGGTACATAGTATGG-GTATCGTGGTACCGTATG
ACGTAAAT-GTATAGAGTCTAGTGGTAC-GTATG
ACGTACATAGTATAGCGACTCGTGGTACCGTATG
```

Arbre trouvé

Échantillonage aléatoire
Avec remise des sites
x 1000

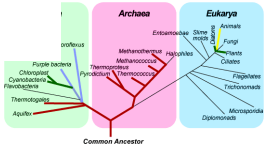
Arbre
« bootstrapé »

1000 Alignements artificiels

```
ATGCCATGCCATTGCTGTACTGGCCCTAGCATGC
ATG-CGTGCGCTTGGTGTACTGGCACTAGGATGC
ATGAAATGACATTGCTGT-CTGG-C-TAGCATGC
ATGCCATGCCCTAGCTGAACTGGCCCTAGCATGC
```

Dénombrement des
occurrences des
branches internes

1000 Arbres artificiels



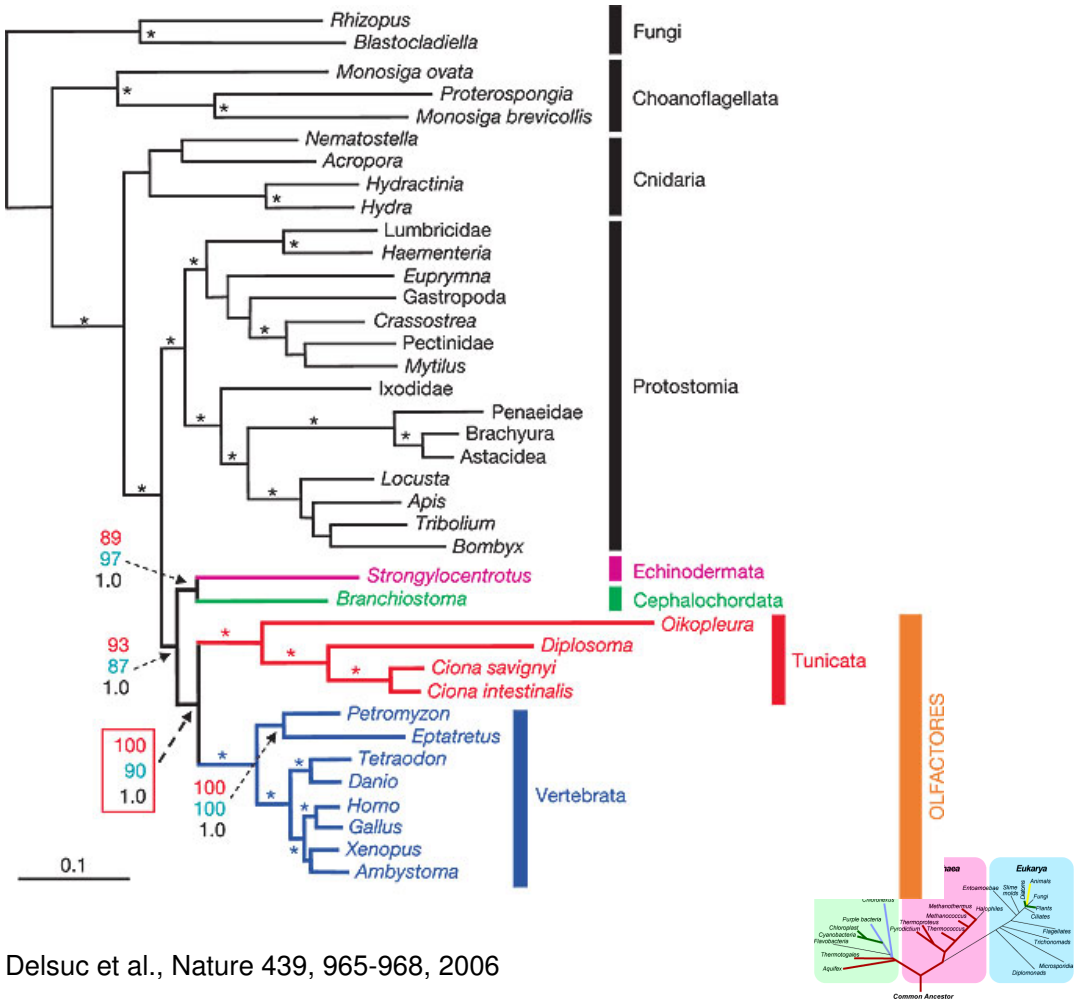
Indices de confiance

Exemple d'arbre avec bootstrap

ML tree obtained from the analysis of 33,800 aligned amino acid positions under a WAG substitution matrix plus a four-category gamma rate correction (= 0.5) using two independent reconstruction algorithms.

Bootstrap proportions obtained after 100 ML (red) and 1,000 MP replicates (blue), as well as bayesian posterior probabilities (black) are shown for selected branches.

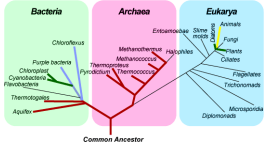
A star indicates that all three values are maximal (100%, 100% and 1.0). Scale bar indicates number of changes per site.



Delsuc et al., Nature 439, 965-968, 2006

2 - Des séquences aux arbres : les étapes

1. Rassembler séquences homologues
2. Aligner ces séquences
3. Reconstruction de l'arbre et méthodes
4. Racinement et interprétation de l'arbre
   - Raciner un arbre
   - Indices de confiance
   - Difficultés d'interprétation
     - Notions d'orthologie et paralogie
     - Différentes vitesses d'évolution
     - Biais compositionels
     - Phénomène de coalescence

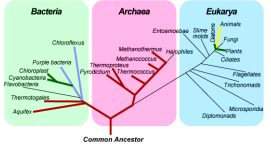


Analyse de l'arbre obtenu

Difficultés d'interprétation

- Alignement erroné
- Saturation, perte du signal phylogénétique : lorsque les séquences homologues que l'on compare ont subi trop de substitutions depuis leur divergence il est impossible de déterminer l'arbre phylogénétique (quelle que soit la méthode utilisée)

- La phylogénie des gènes ne reflète pas la phylogénie des espèces (perte de gènes)
- Vitesses évolutives (Attraction des longues branches)
- Biais compositionels
- Phénomène de coalescence



Difficultés d'interprétation

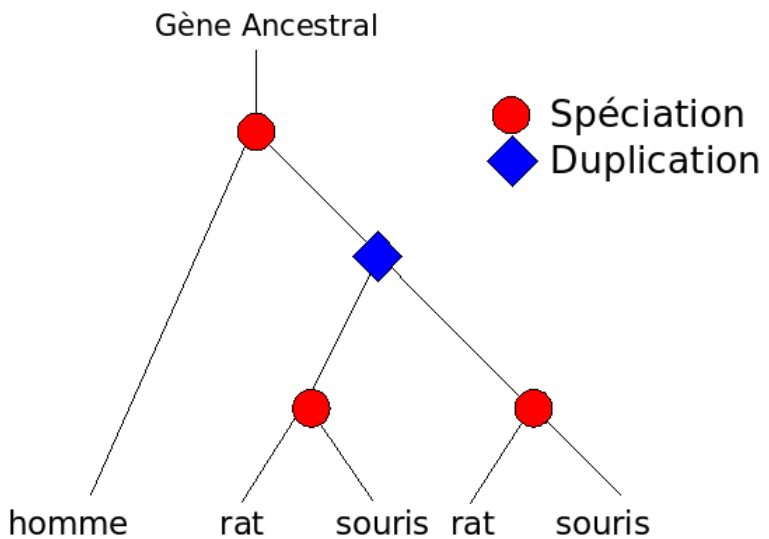
Notions d'orthologie et de paralogie

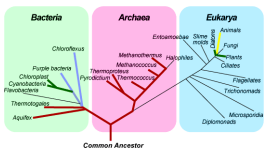
L'histoire évolutive des gènes reflète celle des espèces qui les portent, sauf si :

- Transfert horizontal : transfert de gènes entre espèces
- Duplication de gènes : orthologie/paralogie

Homologie

2 gènes sont homologues s'ils ont un ancêtre commun



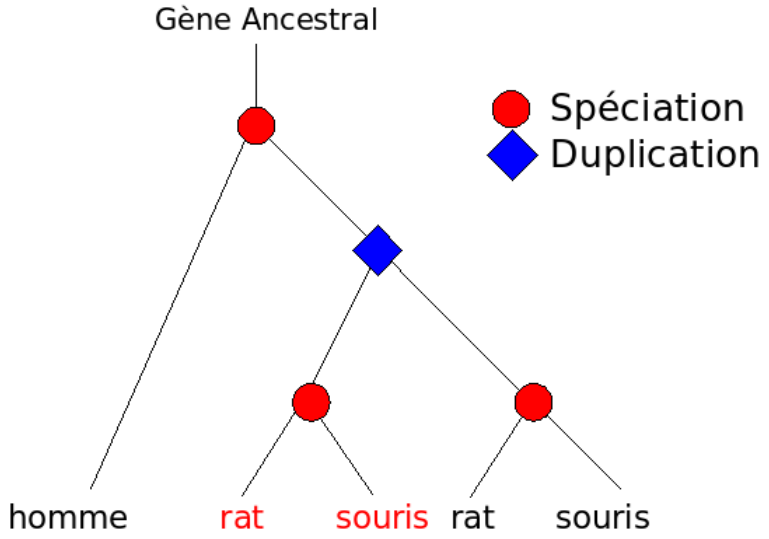


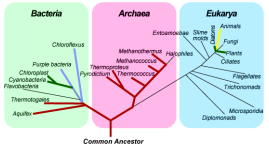
Difficultés d'interprétation

Notions d'orthologie et de paralogie

Orthologie

2 gènes sont orthologues s'ils ont divergé à la suite d'un événement de spéciation



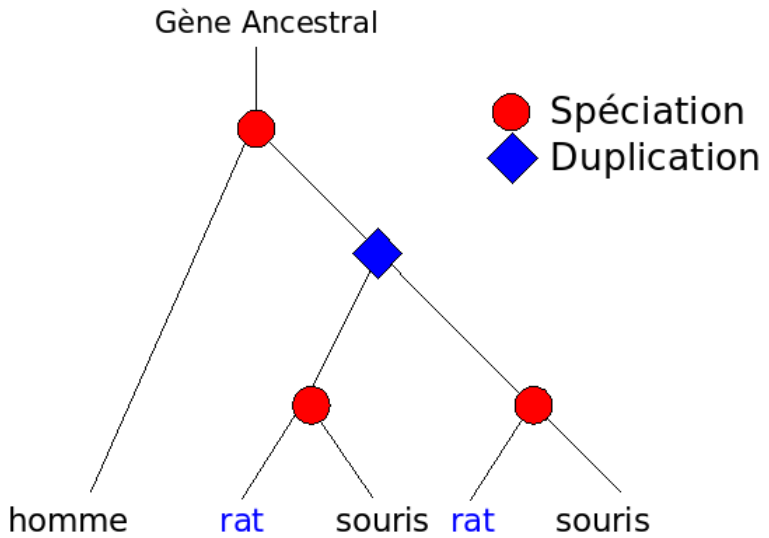


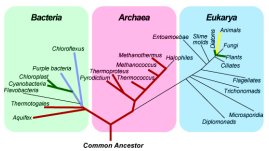
Difficultés d'interprétation

Notions d'orthologie et de paralogie

Paralogie

2 gènes sont paralogues s'ils ont divergé à la suite d'un événement de duplication

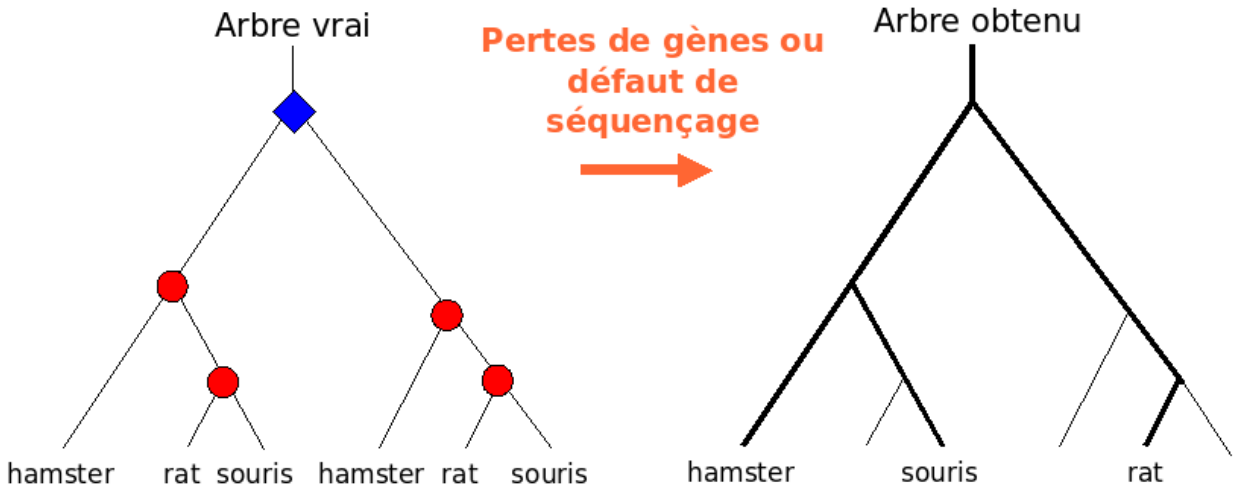


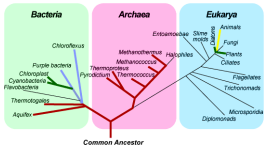


Difficultés d'interprétation

Notions d'orthologie et de paralogie

Erreur liée à la paralogie

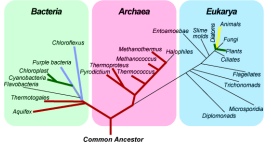




Difficultés d'interprétation

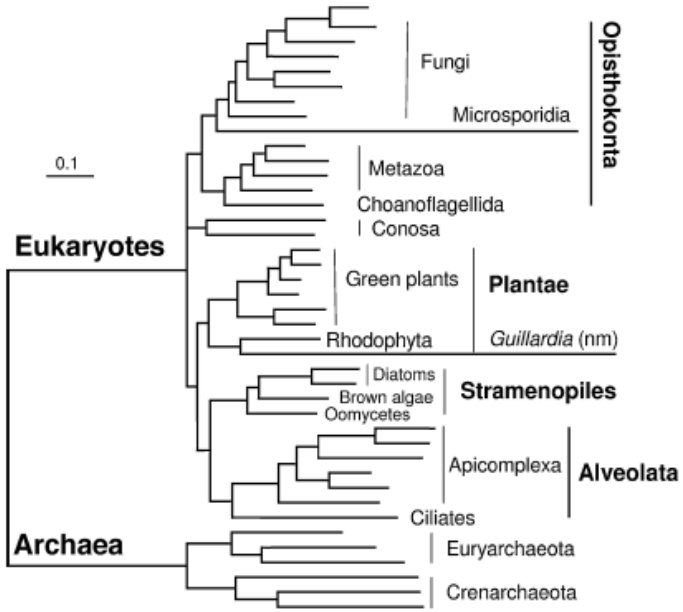
Différentes vitesses d'évolution

- En fonction de la localisation
  - Intron-exon, codant-non codant
  - Nucléaire-mitochondriale
  - 3ème codon
- En fonction de l'expression
- En fonction de l'espèce (attraction des longues branches)

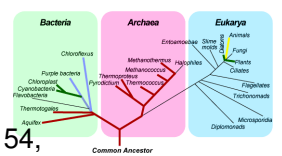


Vitesse d'évolution

Attraction des longues branches



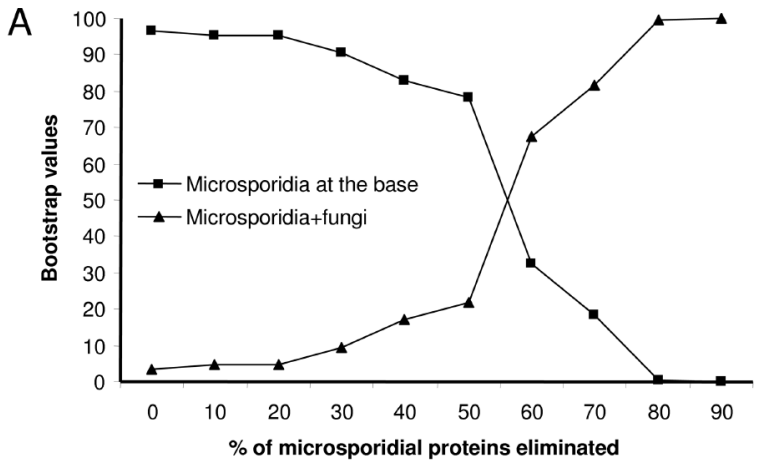
Les séquences de la microsporidie ont un taux d'évolution très important ; de ce fait, l'artefact de l'attraction des longues branches tend à positionner la microsporidie à la base du royaume des Eucaryotes



An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics.
Henner Brinkmann, Mark van der Giezen, Yan Zhou, Gaëtan Poncelin de Raucourt, Hervé Philippe. *Systematic Biology*, Volume 54, Number 5 / October 2005.

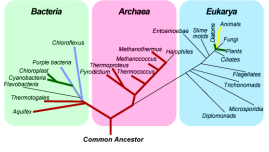
Vitesse d'évolution

Attraction des longues branches



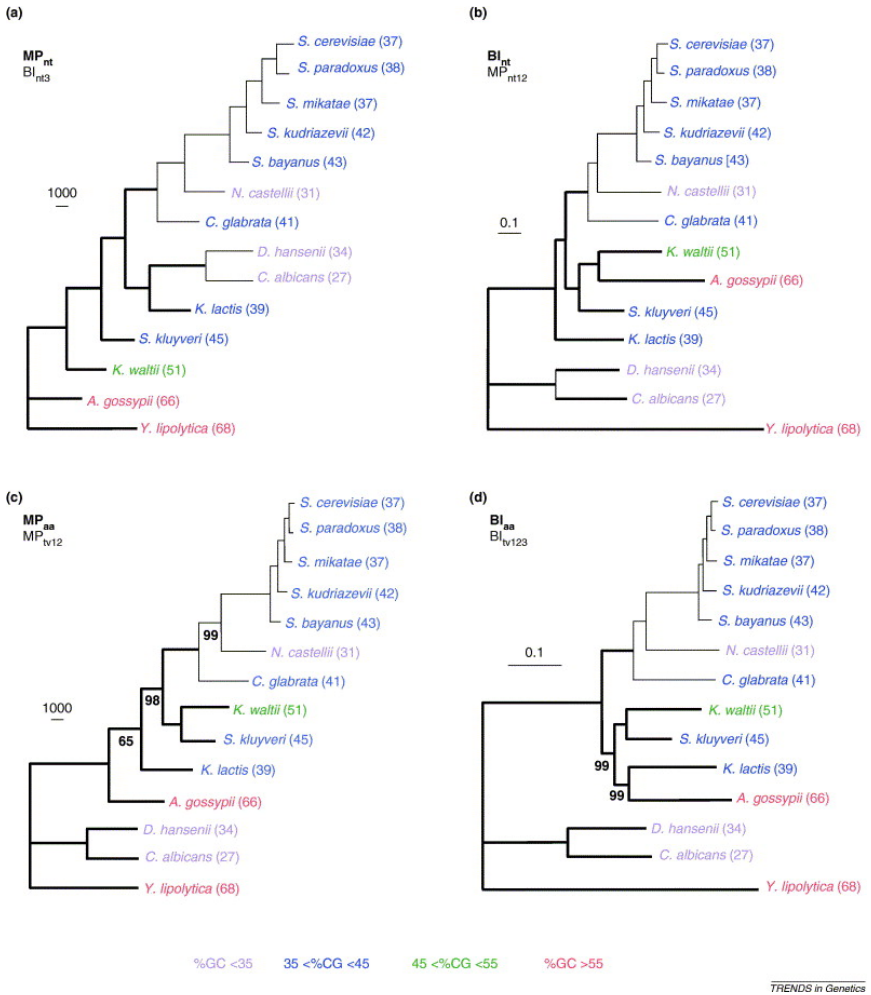
Évolution du soutien (pourcentage de bootstrap)) en faveur de la topologie vraie et de la topologie artefactuelle en fonction du retrait des protéines évoluant le plus vite. Moins il y a de protéines qui évoluent vite, plus la topologie vraie obtient un fort soutien.

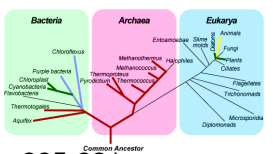
Brinkmann *et al. Systematic Biology*, October 2005.



Difficultés d'interprétation

Biais compositionels



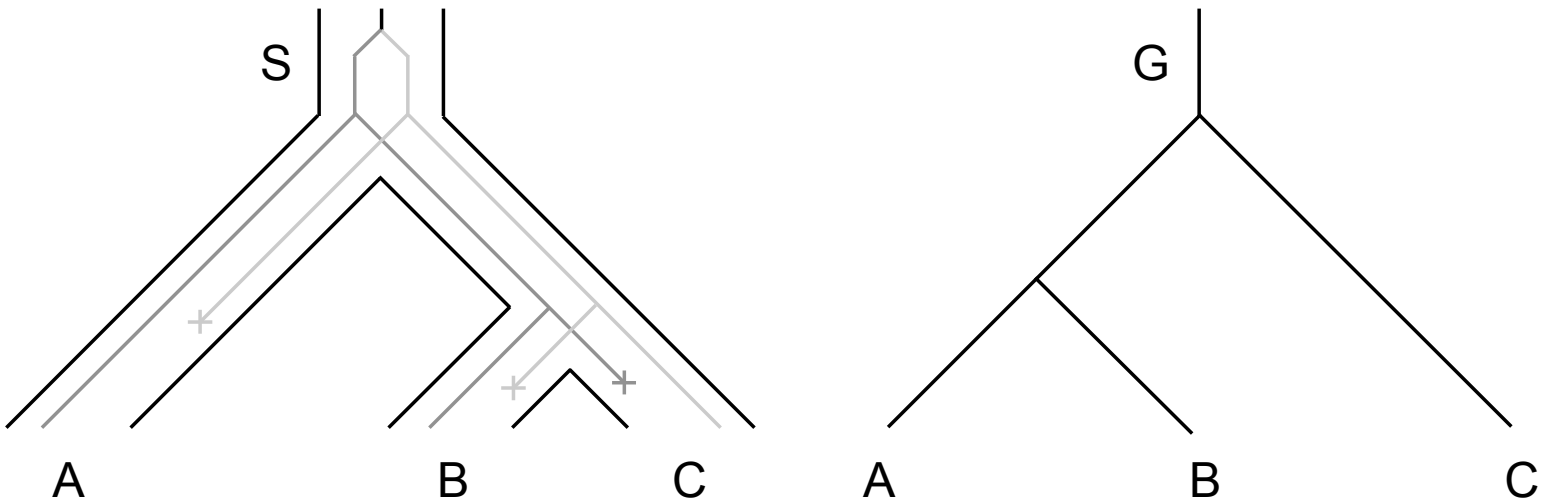


Phylogenomics : the beginning of incongruence ?
Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc and Hervé Philippe, *Trends in Genetics* Volume 22, Issue 4 , April 2006, Pages 225-231

Difficultés d'interprétation

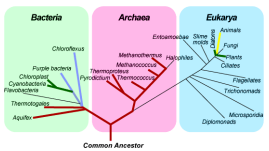
Phénomène de coalescence

Le gène considéré présente deux allèles (gris clair et gris foncé) chez l'ancêtre commun. Les allèles ont disparu préférentiellement selon les descendants.



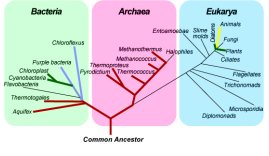
L'histoire des allèles du gène considéré est différente de l'histoire des espèces.

Perrière et Brochier, *Phylogénie Moléculaire*, 2010

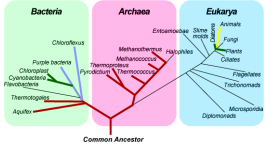


Troisième partie III

Conclusions

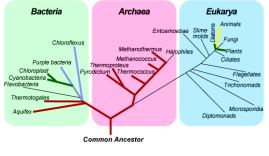


3 - Conclusions

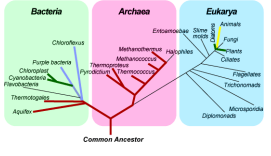


Intérêt des modèles phylogénétiques

- Tiennent compte des substitutions multiples
- Un modèle conduit à faire des hypothèses explicites : on peut donc tester si un jeu de données n'enfreint pas ces hypothèses avant de faire son analyse. Exemples d'hypothèses :
  - les sites évoluent tous à la même vitesse (si enfreint : utiliser une loi gamma)
  - toutes les séquences ont la même composition (si enfreint : utiliser un modèle non homogène)
  - un site évolue à vitesse constante au cours de son histoire (si enfreint : utiliser un modèle hétérotache)
  - il y a eu autant de transitions que de transversions dans l'histoire du jeu de données (si enfreint, utiliser un modèle autre que Jukes et Cantor)

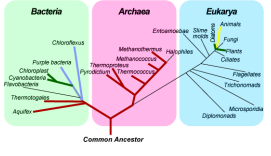


3 - Conclusions

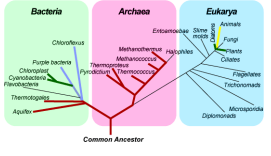


Conclusions

- Une bonne phylogénie nécessite un bon alignement
- Connaître les défauts des méthodes phylogénétiques permet de les détecter, voire de les éviter
- L'existence de méthodes qui font des hypothèses précises (modèles d'évolution) est un atout, pas un inconvénient.



3 - Conclusions



Références

Programmes

- Alignements : Muscle, Clustal
- Éditeur d'alignements visuel et générateur d'arbre : Seaview
- Liste de programmes de phylogénie : `http://evolution.genetics.washington.edu/phylip/software.html`

Ouvrages

- Guy Perrière et Céline Brochier-Armanet : *Phylogénie moléculaire*.